NPS55Bd75062

NAVAL POSTGRADUATE SCHOOL

Monterey, California

CLUSTERING NAVY RATINGS BY LOSS BEHAVIOR

by

R. W. Butterworth
and
P. R. Milch

June 1975

Approved for public release; distribution unlimited

Prepared for:
Naval Personnel Research and Development Center
San Diego, California 92152

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral Linder
Superintendent

Jack R. Borsting
Provost

This work was supported by the Naval Personnel Research and Development Center.

Prepared by:

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS55Bd75062	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Clustering Navy Ratings by Loss Behavior		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) R. W. Butterworth P. R. Milch		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93940		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS PF55.521.010 PO 4-0112
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Personnel Research and Development Center, San Diego, CA 92152		12. REPORT DATE June 1975 13. NUMBER OF PAGES 33
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Rating Cluster Losses

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The enlisted Navy Ratings were clustered by their historical loss behavior, using a hierarchical clustering technique. The immediate application of this clustering technique was to investigate pooling of loss data to improve loss estimation. No significant improvement in loss estimation was found by clustering. Examples of other potential uses for this clustering technique include isolation of groups of ratings to which a common policy regarding loss, reenlistment, etc., may apply.

I. INTRODUCTION

Considerable effort has been spent by the Naval Personnel Research and Development Center (NPRDC), to develop a model that would enable the Navy to forecast future states of the enlisted force structure. This model, entitled FAST, (see [2], [4] and [5]) is a highly comprehensive model that involves acquisitions, losses, and advancements as well as a large number of subcategories of these variables of the Navy personnel force. FAST has been used successfully in the past few years as a long-range planning tool as well as for researching the behavior of the enlisted force. Due to the complexity of the model its operation requires a large amount of data processing and computer time.

In an attempt to increase the flexibility of FAST, this research effort concentrated on a single variable of the personnel force: losses. Since forecasting future losses is one of the major tasks of FAST, it was considered important to attempt to simplify that single aspect of FAST.

II. THE FORECASTING PROBLEM

The enlisted Navy force is organized and managed along the lines of ratings, that is, job skills within the Navy. Consequently, the job of forecasting losses must be done for each rating individually. In addition, losses categorized by length of service and pay grade simultaneously are preferred, so that the effects of projected losses on the force structure can be forecast as well.

When all of the above variables are considered simultaneously, the population of individuals being considered is greatly diminished. For example, while the number of E-5's with 15 years of service may be several hundred, the number of Electronic Technicians who are E-5 with 15 years service is slight.

This problem of sparse data makes the task of accurate forecasting difficult. Procedures for forecasting are all predicated on some statistical stability in people's actions. This stability comes about with large populations of individuals whose reactions are similar. With the small populations that are inherent in sparse data, the consequent lack of statistical stability makes reliable forecasting difficult at best.

To help overcome the problems caused by sparse data, the populations can be recombined to form fewer groups of larger sizes. A natural choice for this combination, or pooling of data, is along the lines of ratings. That is, if ratings which exhibit similar loss behavior statistically are identified and grouped, or clustered together, the resulting clusters can be used in place of ratings to gain some statistical stability. The pooling of data in clusters of ratings is sought only to improve the estimates of loss characteristics and of certain parameters in statistical models. The forecasting of losses for each rating can still be accomplished. This then is one reason for finding clusters of Navy ratings which exhibit similar loss behavior. Other applications of the clustering would be to identify groups of ratings to which common policies regarding loss and retention might be applied. The following sections of this report describe approaches to identifying the clusters and a procedure for estimating their possible effectiveness in improving forecasts.

For the purpose of our analysis, losses were defined to include losses for all reasons, from all pay grades and length of service cells. Actual prediction of losses is more complex, involving many variables, as described in [2] and [4].

III. HIERARCHICAL CLUSTERING

A common technique for clustering is the Hierarchical clustering method. We will give a brief description of the method here, Ref [1] provides more details.

The hierarchical clustering approach groups objects, in our case Navy ratings, into several sets of clusters, each one contained in the previous one. Figure 1 shows a small example of the result for 5 objects.

The tree structure in Figure 1, called a dendrogram, indicates how this procedure formed the groups of clusters. The order shown here is not unlike the groupings which occur in biological taxonomy, where all life forms are grouped, first into species, then into genera, then into families, and so on. This method may appropriately be called numerical taxonomy.

The dendrogram in Figure 1 shows the 5 individual objects being grouped into two groups, objects 1 and 2, and objects 3, 4, and 5. This is the first grouping beyond the base level of 5 singleton groups. A more coarse grouping brings all 5 objects into a single set. The distance scale provides a measure of selectivity in forming the groups. If the "distance" allowed between objects to be clustered

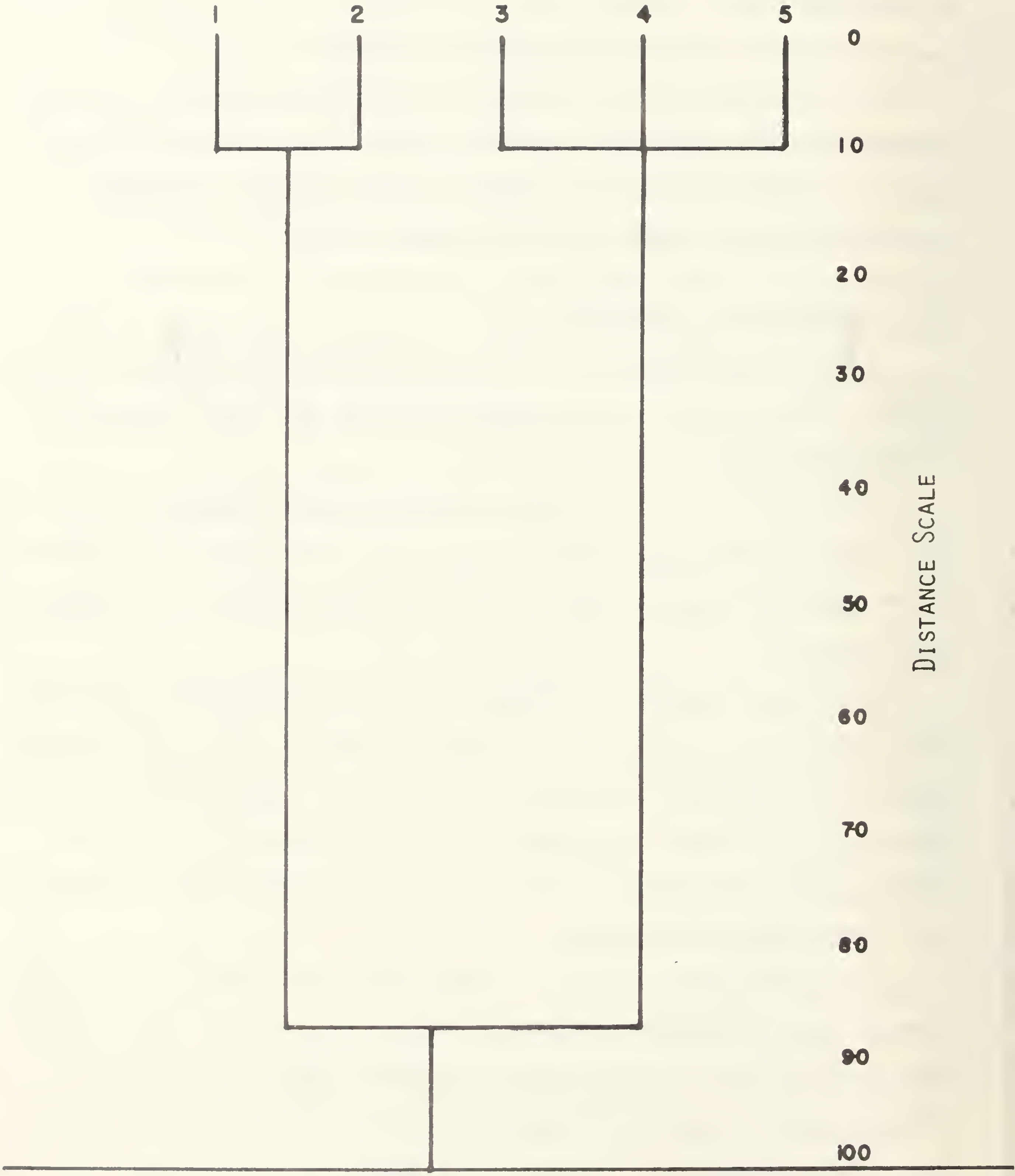


Figure 1: A Dendrogram for Hierarchical Clustering

together is 10, then just two groups are formed. This criterion must be increased to 90 before the first two groups become one, thus indicating that the cluster of two groups is probably natural, while a clustering into one group is probably not. The interpretation of what groupings are natural is somewhat subjective if based only on the dendrogram. As described later, the clusters in this application are evaluated apart from the dendrogram.

In order to produce a dendrogram, a "distance" between each pair of objects must be specified. In this application, the objects are enlisted Navy ratings, and the distance between two ratings should measure the proximity of their loss behavior. The distance function chosen for this purpose is

$$d(k,m) = \left[\sum_{i=1}^{7} \rho^{7-i} (\ell_{i,k} - \ell_{i,m})^2 \right]^{1/2}$$

where

$d(k,m)$ = distance between rating k and m

$\ell_{i,k}$ = loss rate from rating k in year i

$\ell_{i,m}$ = loss rate from rating m in year i

ρ is a parameter, $0<\rho\leq 1$

and years are indexed with 1966 for $i = 1$, 1967 for $i = 2,\ldots,1972$ for $i = 7$. These years are being used simply because they comprise the data base for the research project. The parameter ρ is included to weigh the recent years greater. Thus, two ratings are judged "close" by this criterion if their loss rates are close, especially in recent years. The specific value for the parameter ρ remains to be determined by the methods discussed in a later section.

Once a distance between ratings has been defined, it is necessary to define a distance function between subsets of ratings. This is necessary for the hierarchical clustering algorithm to be defined. While many definitions of distance between subsets are possible, two were investigated and one finally used. The "maximum metric" is defined to be the maximum of all distances between pairs of objects, one choosen from each subset. If C_1 and C_2 are two subsets of ratings, we have

$$d_{max}(C_1,C_2) = \mathrm{Max}\{d(k,m)\,|\,k\varepsilon C_1, m\varepsilon C_2\} \ .$$

The "minimum metric" is analogously defined, with MIN replacing MAX in the above definition.

Under the maximum metric, two subsets of ratings are close only if all ratings are close to each other. The minimum metric only requires that two ratings in the subsets be close, while others may be distant, for the subsets to be close. These two definitions generate strikingly different dendrogram shapes as illustrated later.

IV. CLUSTERING BY CORRELATION

1. Correlating Population Size and Corresponding Loss Rate.

Examination of the data on population sizes and loss rates in various ratings over the years 1966-72 suggested that ratings may be grouped on the basis of whether their population size correlates positively or negatively (and to what extent) with their corresponding loss rate.

For example, it appears that some ratings, such as Quartermaster (200 QM), have their loss rate increase (or decrease) together with their population size over the years 1966-72. At the same time, other ratings, such as Construction Recruit (6000 CR), have their population size and loss rate tend (in most cases) in opposite directions from one year to the next.

The correlation between population size and loss rate was studied for all ratings and "All Navy" over the seven data points, provided by the years 1966-72. In addition to measuring the correlation directly for these data points, rank correlation was also used, since the actual magnitude of the changes in population size seemed both unimportant and incongruous when compared to changes in the loss rate.

Two different rank correlation coefficients were used. These (see [1]) are defined below in terms of the rankings, $P_1, \ldots, P_7$, of the seven population sizes, over the years 1966-72, of a given rating and the rankings $\ell_1, \ldots, \ell_7$ of the seven corresponding loss rates.

(i) Spearman's Rho:

Let $$D_i = P_i - \ell_i \ , \quad i = 1, \ldots, 7$$

be difference in the rankings.

Then $$\rho = 1 - \frac{1}{56} \sum_{i=1}^{7} D_i^2$$

(ii) Kendall's Tau:

Let $$A_{ij} = \begin{cases} +1 & \text{if} \quad (P_i - P_j)(\ell_i - \ell_j) > 0, \\ -1 & \text{if} \quad (P_i - P_j)(\ell_i - \ell_j) < 0 \end{cases} \qquad i,j = 1, \ldots, 7$$

Then $$\tau = \frac{1}{21} \sum_{1 \le i < j \le 7} A_{ij}$$

(iii) Ordinary Correlation Coefficient:

If P_i and ℓ_i denote the actual magnitude of the population sizes and corresponding loss rates respectively of a rating over the years 1966-72, the correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^{7} (P_1 - \bar{P})(\ell_i - \bar{\ell})}{\left[\sum_{i=1}^{7} (P_i - \bar{P})^2 \sum_{i=1}^{7} (\ell_1 - \bar{\ell})^2\right]^{1/2}}$$

where $$\bar{P} = \frac{1}{7} \sum_{i=1}^{7} P_i \text{ and } \bar{\ell} = \frac{1}{7} \sum_{i=1}^{7} \ell_i$$

Each of these correlation coefficients provides a method of clustering of ratings. Kendall's Tau seemed, perhaps, the most accommodating in providing clusters that separate in a somewhat natural way. Thus, three clusters may be formed on the basis of the values of Kendall's Tau:

(i) Ratings with $-1.00 \le \tau \le -0.13$ (Cluster A)

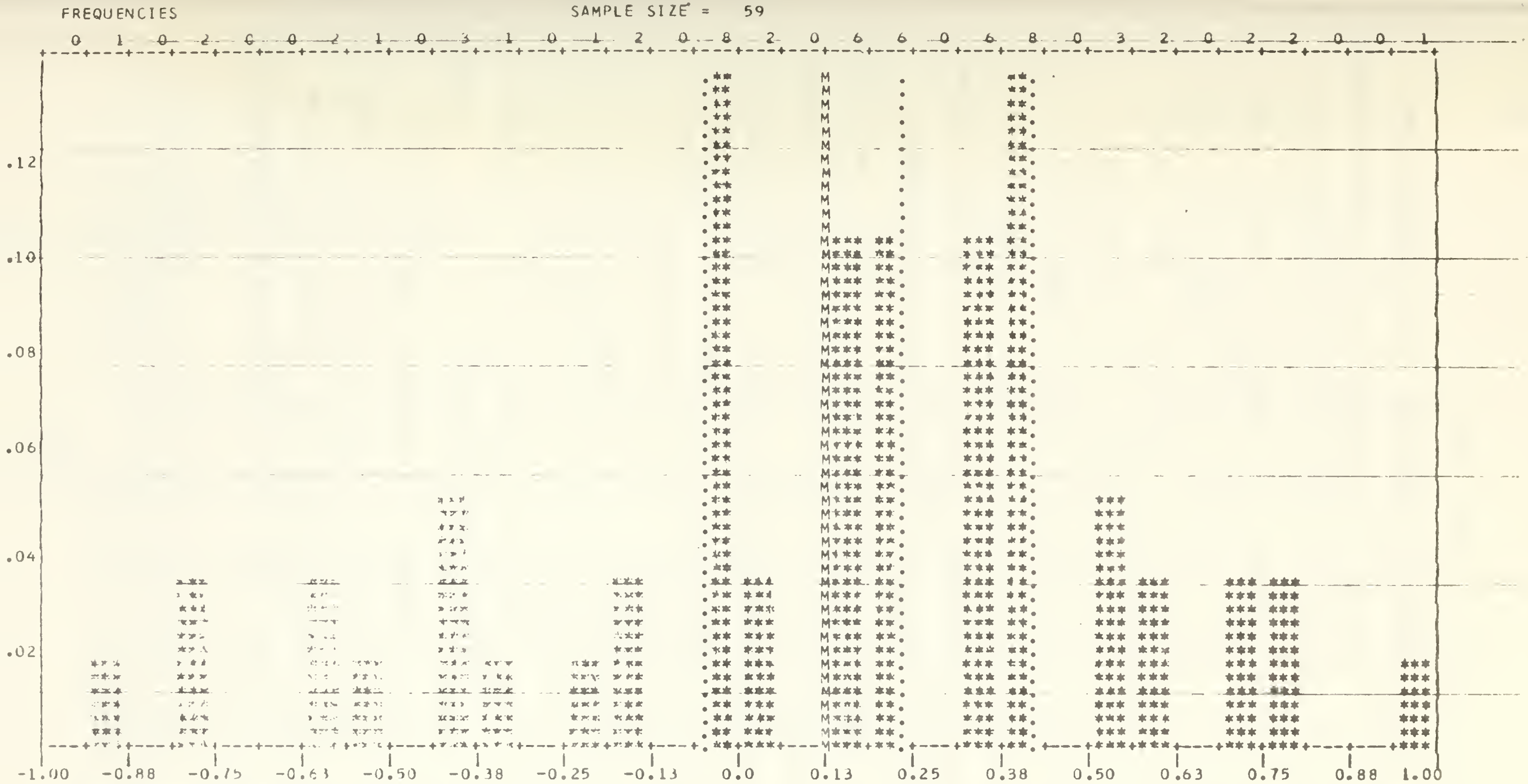
(ii) Ratings with $-0.13 < \tau < +0.50$ (Cluster B)

(iii) Ratings with $+0.50 \le \tau \le +1.00$ (Cluster C)

Table 1 shows a histogram of loss rates for ratings against their τ-values. Each of the three clusters may be broken into further subclusters in various ways based on the loss rates of the ratings in each cluster. Such methods are suggested in the next subsection.

2. Correlating Loss Rates with All Navy Population Size.

If the above procedure for clustering ratings is to be useful it should provide a procedure for forecasting future loss



```
SCALE FIXED FROM -1.CCCOOJE 00 TO  1.000000F 00


CENTRAL TENDENCY            SPREAD                        HIGHER CENTRAL MOMENTS      DISTRIBUTICN

MEAN        1.329257E-01    VARIANCE     1.794637E-01     M3         -4.243908E-02    MINIMUM                     -9.047619E-01
MEDIAN      2.380952E-01    STD DEV      4.236316E-01     M4          9.480608E-02    .10 QUANTILE                -6.190476E-01
TRIMEAN     2.142857E-01    COEF VAR     3.186883E 00     SKEWNESS   -5.582147E-01    .25 QUANTILE (HINGE)        -4.761904E-02
MIDMEAN     1.858678E-01    MEAN DEV     3.274415E-01     KURTOSIS   -5.637360E-02    .50 QUANTILE (MEDIAN)        2.380952E-01
MIDRANGE    4.761907E-02    RANGE        1.904761E 00     BETA1      -4.030555E-02    .75 QUANTILE (HINGE)         4.285714E-01
                            MIDSPREAD    4.761904E-01     BETA2       9.157163E-02    .90 QUANTILE                 6.190476E-01
                                                                                      MAXIMUM                      1.000000E 00
```

TABLE I: KENDAL'S TAU

rates through the use of clusters. Since the above clusters are obtained by correlating loss rates of ratings with the corresponding population sizes, one would have to have reasonably accurate estimates of future population sizes in each rating in order to forecast corresponding loss rates (and then actual losses). It seems unlikely that such estimates would be available for each rating and certainly not several years in advance. If good estimates of population sizes will be available for future years at all it will be for "All Navy" only. For that reason, it appears desirable to correlate loss rates of ratings with "All Navy" population size. The three correlation coefficients defined above are again relevant with the only change that $P_1,...,P_7$ now denote the "All Navy" population sizes, or their rankings, over the years 1966-72. Table 2 presents the lists of ratings in three clusters formed on the basis of Kendall's Tau. The three clusters are:

(i) Ratings with $-1.00 \leq \tau \leq -0.15$ (Cluster A)

(ii) Ratings with $-0.15 < \tau \leq +0.25$ (Cluster B)

(iii) Ratings with $+0.25 < \tau \leq +1.00$ (Cluster C)

All three of these clusters may be considered too big and in any case loss rates of ratings within each cluster vary widely. Since clusters are envisioned as groups of ratings of like loss rates it is necessary to break each of the above clusters into further subclusters. (The same remark applies when clustering is accomplished based on correlating each loss rate with its own population size.)

Further subclusters may be formed by selecting one of several candidate statistics, such as:

LOSS RATES OF CLUSTER A RATINGS

			1966	1967	1968	1969	1970	1971	1972	*TAU
3600	SR	SEAMAN RECRUIT	19.90	16.94	19.80	27.86	29.11	38.70	37.22	-0.43
300	OS	OPERATIONS SPECIALIST	21.92	28.81	29.44	29.25	30.87	31.17	31.26	-0.43
7800	AR	AIRMAN RECRUIT	19.93	17.19	13.26	16.43	20.50	31.16	32.02	-0.43
1100	IM	INSTRUMENTMAN	13.41	22.00	26.77	29.93	33.02	36.01	39.04	-0.33
7500	AS	AV. SUPPORT EQUIP. TECH.(4)	0.0	0.0	15.06	13.15	25.88	26.12	20.01	-0.29
5000	FR	FIREMAN RECURIT	14.86	13.64	16.92	22.02	28.99	34.08	28.64	-0.24
3200	DM	ILLUSTRATOR CRAFTSMAN	22.39	25.38	26.59	28.60	40.34	42.49	30.52	-0.14
8500	SD	STEWARD	9.80	8.21	6.44	4.67	5.40	7.33	7.12	-0.14
900	MN	MINEMAN	9.18	17.67	13.34	26.47	23.26	30.54	25.97	-0.14
6200	AD	AVIATION MACHINSTS MATE(3)	17.47	22.64	22.87	17.96	24.62	26.59	24.02	-0.14
7700	PT	PHOTOGRAPHIC INTELLIGENCE	13.65	18.06	20.57	18.81	36.27	37.14	20.15	-0.14
6000	CP	CONSTRUCTION RECRUIT	8.56	10.71	18.12	20.46	38.15	39.28	32.35	-0.14
8300	DT	DENTAL TECHNICIAN	15.75	25.10	23.36	22.00	30.21	26.92	30.33	-0.14

LOSS RATES OF CLUSTER B RATINGS

			1966	1967	1968	1969	1970	1971	1972	*TAU
602	GMT	GUNNERS MATE (TECHNICIAN)	14.16	21.54	18.35	15.68	21.98	19.75	23.67	-0.05
0		ALL NAVY	18.00	20.94	25.69	29.46	34.15	32.38	30.86	-0.05
1010	DS	DATA SYSTEMS TECHNICIAN	20.70	18.16	11.94	9.52	13.23	12.27	13.75	-0.05
2490	SH	SHIPS SERVICEMAN	16.43	27.94	28.94	33.13	37.93	34.24	30.56	0.05
7600	PH	PHOTOGRAPHERS MATE	19.04	24.23	26.44	21.84	32.04	28.52	25.20	0.05
6900	AM	AVIATION STRUCTURAL MECH(4)	15.31	19.04	21.95	16.59	25.35	23.59	20.95	0.05
6800	AE	AVIATION ELECTRICIANS MATE	17.84	20.01	21.54	18.99	25.42	23.73	20.91	0.05
8000	HM	HOSPITAL CORPSMAN	19.75	21.76	19.67	19.80	32.98	24.98	22.95	0.05
3800	EN	ENGINEMAN	18.16	28.96	27.14	27.31	36.99	30.23	32.65	0.05
4600	PM	PATTERNMAKER	17.50	23.43	33.88	19.81	33.63	30.73	25.00	0.05
7300	AK	AVIATION STOREKEEPER	19.72	21.48	21.70	22.28	30.48	32.02	19.80	0.14
3900	MR	MACHINERY REPAIRMAN	19.74	30.36	30.66	29.94	36.93	29.09	33.53	0.14
7000	PR	AIRCREW SURVIVAL EQUIPMAN	15.57	20.03	16.50	16.37	22.88	22.63	19.81	0.14
1701	LN	LEGALMAN	12.35	12.52	19.31	32.86	46.86	32.42	30.44	0.14
500	TM	TORPEDOMANS MATE	12.77	22.77	21.97	21.19	25.77	21.59	23.32	0.14
6500	AO	AVIATION ORDANCEMAN	18.24	22.77	21.29	20.23	29.05	23.53	22.56	0.14
2700	PC	POSTAL CLERK	24.98	37.05	38.91	44.08	53.77	42.12	40.23	0.14
3700	MM	MACHINISTS MATE	17.61	24.34	25.48	26.63	29.19	25.17	25.90	0.14
2290	CS	COMMISSARYMAN	14.44	23.04	22.67	24.92	29.64	24.28	24.80	0.14
2600	JO	JOURNALIST	25.88	34.21	32.02	33.94	41.72	41.68	38.09	0.14
3300	MU	MUSICIAN	19.27	21.63	13.89	14.29	32.56	24.45	18.17	0.14
600	GM	GUNNERS MATES(3)	17.67	25.76	25.38	27.27	38.39	28.09	26.11	0.14
3100	LI	LITHOGRAPHER	30.67	37.89	34.43	33.91	47.55	34.43	39.87	0.14
6600	AC	AIR CONTROLMAN	14.02	21.64	19.26	17.44	26.59	25.14	21.59	0.14
4700	ML	MOULDER	12.65	26.25	24.89	29.91	26.22	24.02	28.51	0.24
4200	IC	INTERIOR COMMUNICATION ELEC.	18.79	27.64	27.44	28.95	37.10	24.81	29.00	0.24
1200	OM	OPTICALMAN	16.53	25.26	26.01	24.63	24.70	21.29	24.87	0.24
100	BM	BOATSWAINS MATE	17.77	29.55	33.36	37.96	42.57	33.46	30.18	0.24
810	MT	MISSILE TECHNICIAN	4.90	7.76	11.91	17.94	17.71	10.85	10.42	0.24

LOSS RATES OF CLUSTER C RATINGS

			1966	1967	1968	1969	1970	1971	1972	*TAU
200	QM	QUARTERMASTER	22.85	31.67	28.06	34.12	36.17	32.76	31.19	0.33
1900	DP	DATA PROCESSING TECHNICIAN	21.02	25.47	22.55	24.75	35.39	23.75	25.36	0.33
7100	AG	AEROGRAPHERS MATE	15.65	24.15	21.38	21.10	27.74	25.47	20.34	0.33
7400	AZ	AV. MAINT. ADMINISTRATION	27.28	32.16	30.37	29.4[illegible]	39.06	40.72	24.55	0.33
2000	SK	STOREKEEPER	17.20	25.25	26.87	28.74	35.74	27.48	24.93	0.33
1500	RM	RADIOMAN	17.79	22.99	22.96	26.45	28.59	22.95	24.24	0.33
4100	EM	ELECTRICIANS MATE	17.78	27.10	27.12	26.81	30.51	23.66	27.08	0.33
4000	BT	BOILERMAN(2)	20.33	30.38	27.72	31.64	32.95	26.37	31.01	0.33
6700	AB	AVIATION BOATSWAINS MATE(4)	21.89	32.68	29.43	27.69	37.20	35.50	22.43	0.33
250	SM	SIGNALMAN	19.35	27.58	27.13	29.81	31.54	25.80	27.36	0.43
2100	DK	DISBURSING CLERK	18.33	26.76	29.53	30.99	30.54	26.60	26.37	0.43
7200	TD	TRADEVMAN	11.02	15.40	19.81	19.04	25.05	13.66	12.23	0.43
1800	PN	PERSONNELMAN	20.31	25.19	25.91	30.20	31.86	25.61	22.19	0.43
4500	DC	DAMAGE CONTROL	20.41	28.94	24.61	32.27	41.86	29.09	17.69	0.52
400	ST	SONAR TECHNICIANS(3)	17.01	23.52	20.83	24.32	27.75	15.73	18.18	0.62
1000	ET	ELECTRONICS TECHNICIANS(3)	18.34	23.74	24.01	24.21	25.60	13.97	13.69	0.71
800	FT	FIRE CONTROL TECHNICIANS(4)	19.12	26.18	22.26	25.25	27.72	18.55	16.01	0.90

Table 2

(i) The mean loss rate of ratings over the seven years;

(ii) The median loss rate of ratings over the seven years;

(iii) The mean or median loss rate of ratings over the last three years only;

(iv) The loss rate of ratings of the last year only.

For demonstration purposes, one of these statistics, namely the median loss rate of ratings over the three years 1970-72, was selected. Figure 2 shows each of the ratings (and "All Navy") represented by its median loss rate over the years 1970-72. The three clusters referred to above are separated in the graph. The graph itself suggests further subclusters based on the size of the loss rates. For example, Cluster A may be grouped in four subclusters based on the median loss rate $\ell_i^{(m)}$ of (ii):

(a) Ratings in Cluster A with $0\% \leq \ell_i^{(m)} \leq 20\%$ (A_1)

(b) Ratings in Cluster A with $20\% < \ell_i^{(m)} \leq 27\%$ (A_2)

(c) Ratings in Cluster A with $27\% < \ell_i^{(m)} \leq 33\%$ (A_3)

(d) Ratings in Cluster A with $33\% < \ell_i^{(m)} \leq 100\%$ (A_4)

Similar subclusters may be formed within Clusters B and C. These are indicated in Figure 2 by vertical lines drawn as boundaries between neighboring subclusters.

Shortcomings of this method are that it is quite "ad hoc" in selecting the boundaries between clusters and subclusters. Also, since at the start clusters are formed based on values of the correlation coefficients, ratings of similar losses may be found in separate clusters. Thus, e.g. many ratings in Cluster C have loss rates closer to those of some ratings in Cluster B than those

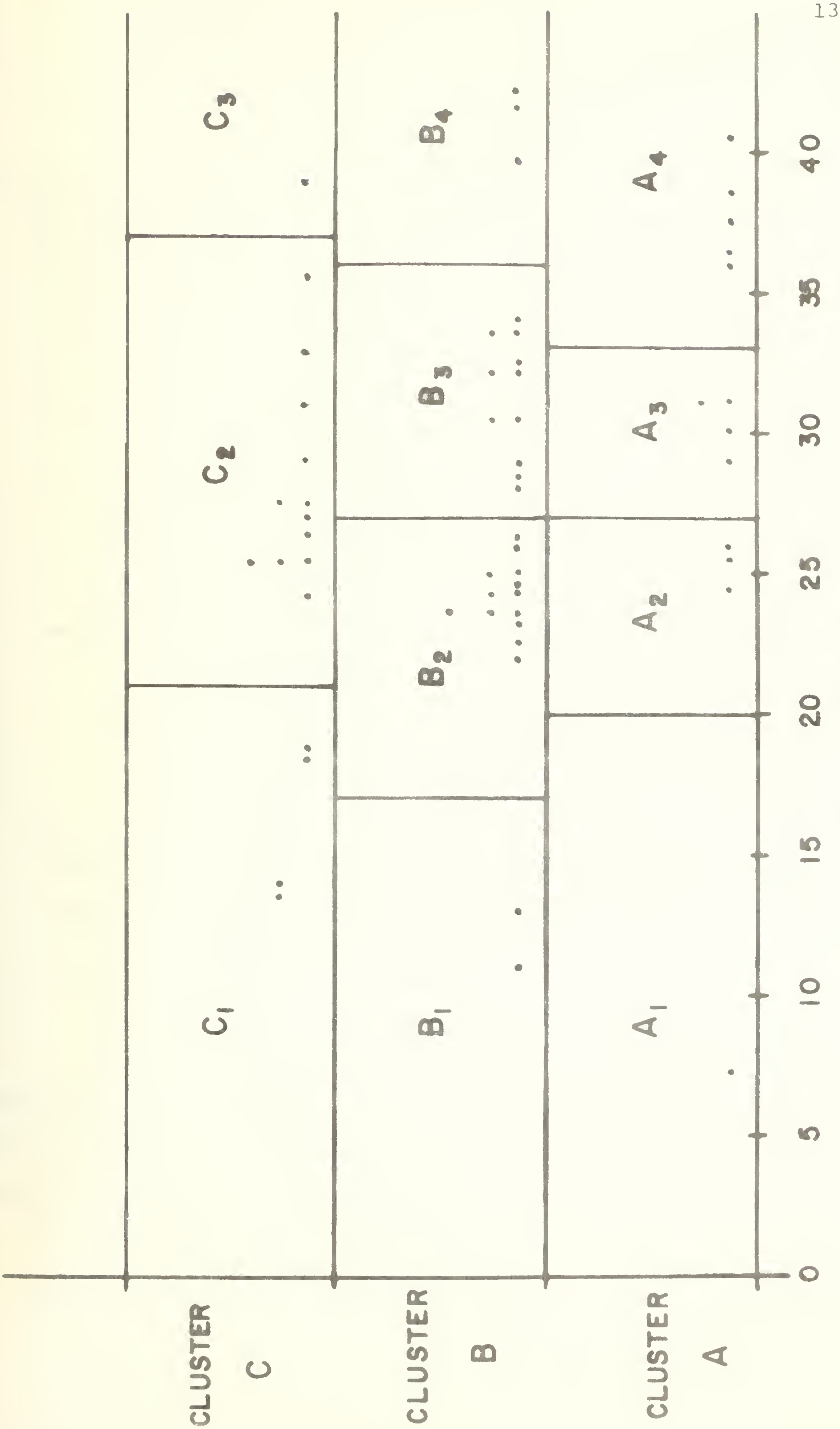


Figure 2: LOSS RATES IN %

of ratings in their own subcluster. This may be regarded as a disadvantage if one considered it an overriding necessity to cluster by like loss rates. On the other hand, ratings with similar loss rates may be placed in different clusters, because these loss rates may be tending in opposite directions over the years. It may be desirable in such cases to group such ratings separately despite their like loss rates.

Because of the ad hoc nature of this clustering method it was not used in the rest of this research effort.

V. EVALUATION OF HIERARCHICAL CLUSTERS

The methods described above lead to various clusterings or partitions of the enlisted ratings. In this section, we describe how any such partition was evaluated.

Let the set of enlisted ratings be designated S, where

$$S = \{1,2,...,N\}$$

and N is the number of ratings being considered. In our case, $N = 71$ ratings. The total number of individual ratings is about 130, however some of the 130 are service ratings which support a general rating. In these instances, several service ratings contain men specializing in a similar area, usually at the middle paygrades such as E4 to E6 or E7. A single general rating associated with these service ratings contains all men at the pay grades beyond those of the service rating, in the common area. The general rating then contains the foremen and line managers for the men in the service ratings. When this occured, all the service ratings and its associated general rating were combined into a pseudo rating for the analysis. This avoided having ratings with only a few pay grades. The common technical skill areas of these ratings made their prior combination seem natural, and reduced the number of ratings analyzed to 71. A few recent ratings with no history in our data base were left out, as they were a special case and quite few in number. The following table shows the definition of ratings used for the study, with the actual rating codes included in each of our ratings.

With the ratings as defined above, a partition or clustering of S is a set of subsets C_k of S for which

$$C_k \cap C_j = 0 \quad \text{if} \quad k \neq j$$

$$\bigcup_k C_k = S$$

If there are m subsets $C_k (k=1,...,m)$, the partition is said to be of size m. Many partitions, suggested primarily by the hierarchical clustering method, were evaluated by a method described below.

This research investigation was conducted for the express purpose of finding out if the prediction of losses by forecasting loss rates could be improved when data was pooled among ratings in clusters, for some systematically well-defined clustering. The approach was to forecast losses by a method approximating the one actually used, and for which the clustering was originally intended. The forecasting was done for the year 1973 (fiscal year), using

RATINGS USED IN THE STUDY

Index in S	Name	Rating Codes
1	Boatswains Mate	100
2	Quartermaster	200
3	Signalman	250
4	Operations Specialist	300
5	Sonar Technicians	400, 401, 404
6	Torpedomans Mate	500
7	Gunners Mates	600, 601, 604
8	Gunners Mate Technician	602
9	Fire Control Technicians	800, 801, 802, 803
10	Missile Technician	810
11	Mineman	900
12	Electronics Technicians	1000, 1001, 1002
13	Data Systems Technician	1010
14	Instrumentman	1100
15	Opticalman	1200
16	Radioman	1500
17	Communication Technicians	1600, 1611, 1622, 1633, 1644, 1655, 1666
18	Yeoman	1700
19	Legalman	1701
20	Personnelman	1800
21	Data Processing Technician	1900
22	Storekeeper	2000
23	Disbursing Clerk	2100
24	Commissaryman	2290
25	Ships Serviceman	2490
26	Journalist	2600
27	Postal Clerk	2700
28	Lithographer	3100
29	Illustrator Draftsman	3200
30	Musician	3300

Index in S	Name	Rating Codes
31	Seaman Recruit	3600
32	Machinists Mate	3700
33	Engineman	3800
34	Machinery Repairman	3900
35	Boilerman	4000,4020
36	Electricians Mate	4100
37	Interior Communication Elec.	4200
38	Hull Technicians	4300, 4410, 4411, 4412
39	Damage Control	4500
40	Patternmaker	4600
41	Moulder	4700
42	Fireman Recruit	5000
43	Engineering Aid	5100, 5101, 5102
44	Construction Electrician	5300, -1, -2, -3, -4, -5, -6
45	Equipment Operator	5410, 5411, 5412
46	Construction Mechanic	5500, 5503, 5504
47	Builder	5600, 5601, 5602, 5603
48	Steel Worker	5700, 5703, 5704
49	Utilitiesman	5800, 5801, 5802, 5803, 5804
50	Construction Recruit	6000
51	Aviation Machinists Mate	6200, 6205, 6206
52	Aviation Electronics Technician	6300, 6304, 6306, 6307
53	Aviation Antisub Warfare Technician	6310
54	Aviation Ordanceman	6500
55	Aviation Fire Control Technician	6520, 6521, 6522
56	Air Controlman	6600
57	Aviation Boatswains Mate	6700, 6704, 6705, 6706
58	Aviation Electricians Mate	6800
59	Aviation Structural Mechanic	6900, 6901, 6902, 6903
60	Aircrew Survival Equipman	7000

Index in S	Names	Rating Codes
61	Aerographers Mate	7100
62	Tradevman	7200
63	Aviation Storekeeper	7300
64	Aviation Maintenance Admin.	7400
65	Aviation Support Equip. Technician	7500, 7501, 7502, 7503
66	Photographers Mate	7600
67	Photographic Intelligence	7700
68	Airman Recruit	7800
69	Hospital Corpsman	8000
70	Dental Technician	8300
71	Steward	8500

TABLE 3

data in the years 1966-72. Then, the predicted losses were compared to the actual losses in 1973. The prediction scheme was not detailed enough to be used for actually forecasting losses, and was only intended to be an evaluation of clustering. If clustering is to improve significantly the forecasting (by any means), then it should improve forecasting by the elementary prediction scheme given below.

To evaluate any clustering or partition C_k, $k=1,\dots,m$, the following approach was used. First, a projection of total losses was made for each individual rating by projecting the loss rate, i.e., the proportion of those on board at the year's start who would be lost over the year. Let

$I_{i,j}$ = Inventory (of men) at the beginning of year i, in rating j.

$L_{i,j}$ = Losses during year i from rating j.

where the indices are,

$i = 1,2,\dots,7$ for years 1966, 1967,...,1972 respectively, and

$j = 1,2,\dots,N$.

The estimated loss rate in 1973 for rating j, denoted $\hat{\ell}_j$, was obtained from a weighted average of the actual loss rates in prior years. Specifically,

$$\hat{\ell}_j = \frac{\sum_{i=1}^{7} \alpha^{7-i}(L_{i,j} \div I_{i,j})}{\sum_{i=1}^{7} \alpha^{7-i}} ,$$

where α is a fixed weighting factor, $0 < \alpha \leq 1$. This estimated loss rate was applied to the 1973 inventory I_j, yielding

$$\hat{L}_j = \hat{\ell}_j \cdot I_j$$

as the estimated loss from rating j in 1973, using no clustering.

The same prediction scheme was used with clustering, and both predictions were compared to the actual loss. To estimate the loss rate with clusters, let C_k $k= 1,2,\ldots,m$ be the partition of the ratings being considered. Then, pooling data over clusters gives the formula for the common estimated loss rate of ratings in cluster C_k:

$$\tilde{\ell}_j = \frac{\sum_{i=1}^{7} \alpha^{7-i} \left(\sum_{j \in C_k} L_{i,j} \div \sum_{j \in C_k} I_{i,j} \right)}{\sum_{i=1}^{7} \alpha^{7-i}}$$

for every $j \in C_k$. Then the estimated loss is

$$\tilde{L}_j = \tilde{\ell}_j \cdot I_j$$

It should be emphasized again that the prediction scheme used here is not intended to be the best available for the data at hand. Our purpose is only to evaluate the clustering, by comparing loss predictions with and without clustering, using the same prediction scheme in both instances.

VI. RESULTS OF CLUSTERING EXPERIMENT

1. Dendrograms.

Using the distance function defined in Chapter III, two dendrograms were drawn for each of several values of the weighting factor ρ. The two dendrograms correspond to the maximum and the minimum metrics, respectively, between clusters as defined in Chapter III. Figures 3 and 4 show examples of dendrograms with the minimum and maximum metric respectively. An undersirable feature of all dendrograms with the minimum metric is, as can be seen in Figure 3, that separation into clusters does not occur until sets are at a fairly close "distance" to each other. For example, in Figure 4, although two clusters form at a "distance" of 15.60, the next separation into (three) clusters occurs at a "distance" of 3.12. Further separations occur at very short intervals, at "distance" values 2.25, 1.692, 1.688, etc. This makes it rather difficult to decide on the number of clusters to be used. In contrast, Figure 4 shows a typical dendrogram with the maximum metric. Here separations into clusters occur quite gradually at least until about ten clusters have formed. Separation into two, three, four, etc., clusters occur at the "distance" values 48.7, 29.9, 18.2, 14.3, 9.4, 7.6, etc. This provides more justification to choose e.g., four clusters rather than three or five. In choosing the appropriate number of clusters one must consider that, while too many clusters would defeat the purpose of clustering, too few clusters would result in a prediction method that is too crude. For this reason the proper choice is probably be somewhere between three and ten clusters.

2. Evaluation of Clustering.

In order to evaluate the effectiveness of clustering, the prediction scheme described in Chapter V was devised. According to this scheme, two estimates, $\tilde{L}_j$ and $\hat{L}_j$, were computed as predictions with and without clustering for the losses in 1973 from

FIGURE 3

DENDROGRAM WITH MINIMUM METRIC, ρ = 0.1

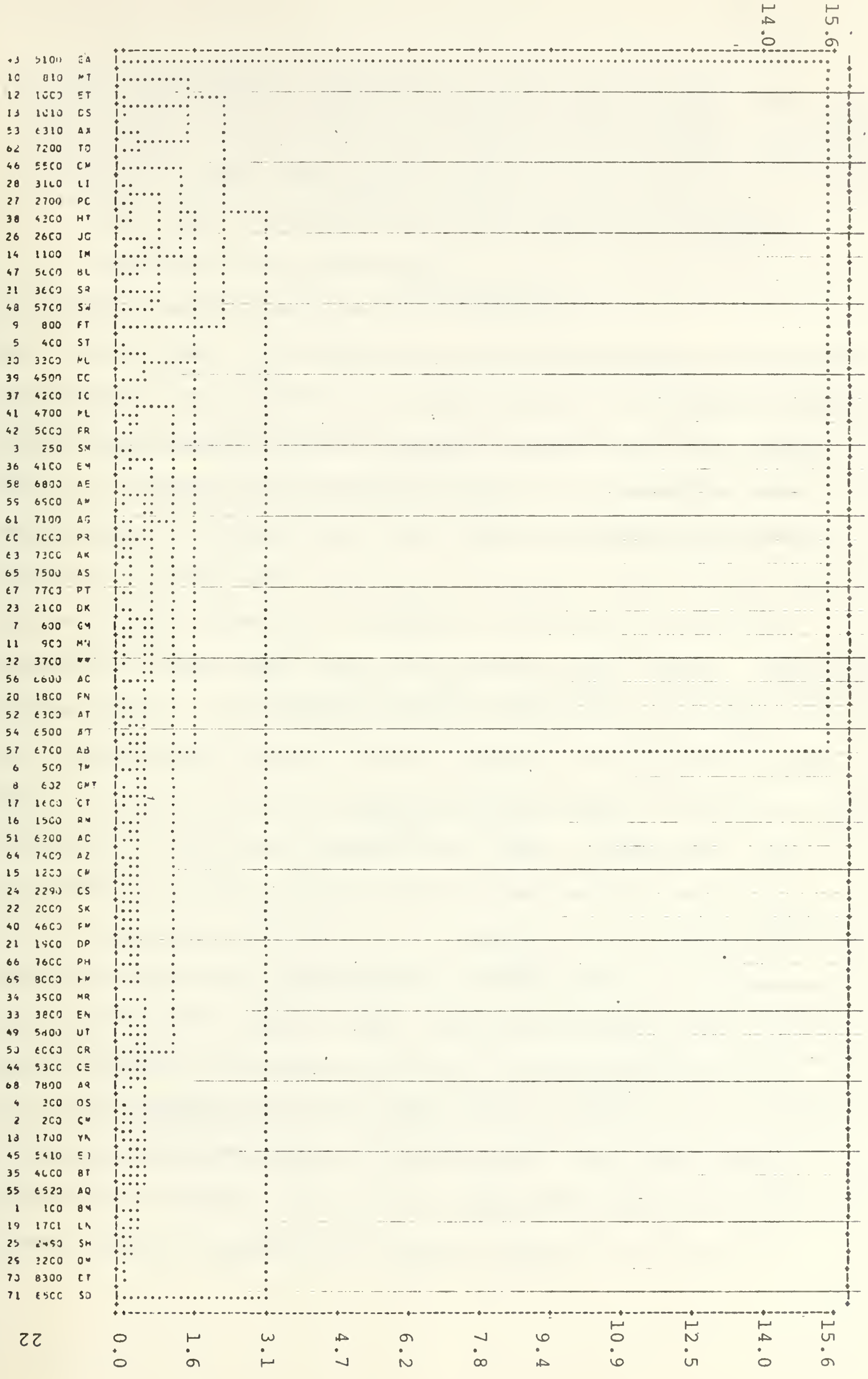
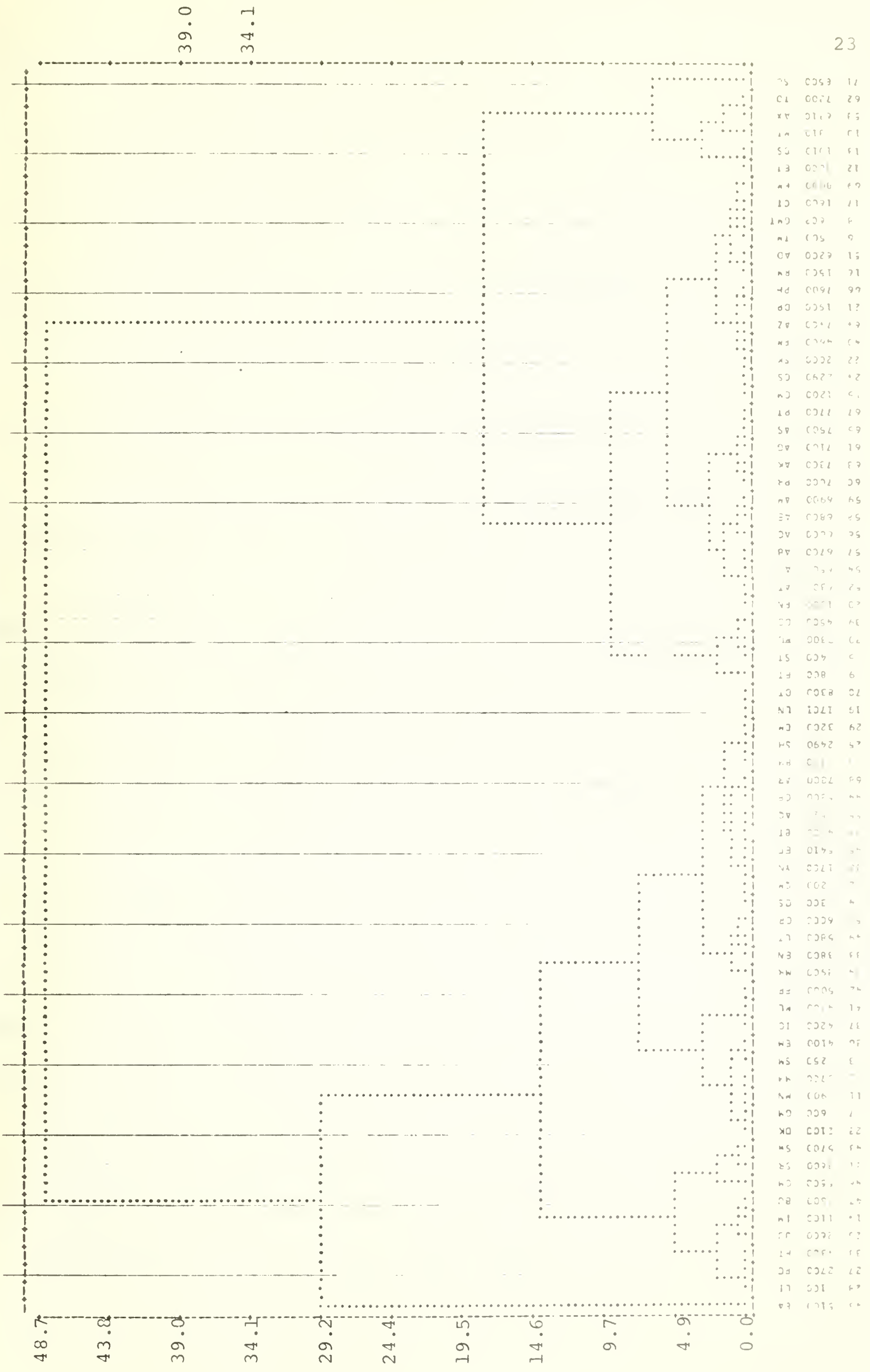


FIGURE 4

DENDROGRAM WITH MAXIMUM METRIC, ρ = 0.1



Rating j. When the 1973 data on losses became available, the actual losses, L_j, from Rating j became known. Histograms were then prepared for the following expressions:

(i) $L_j - \hat{L}_j$ = error in prediction without clustering.

(ii) $L_j - \tilde{L}_j$ = error in prediction with clustering.

(iii) $|L_j - \hat{L}_j| - |L_j - \tilde{L}_j|$ = difference in absolute errors without and with clustering.

(iv) $(L_j - \hat{L}_j) \div L_j$ = normalized error in prediction without clustering

(v) $(L_j - \tilde{L}_j) \div L_j$ = normalized error in prediction with clustering

(vi) $(|L_j - \hat{L}_j| - |L_j - \tilde{L}_j|) \div L_j$ = difference in absolute normalized errors without and with clustering.

The histograms were specifically examined for cases where the number of clusters was 3, 5, 7, 10, 15 and 20.

The proper choice of value for ρ, the parameter used to weight past years according to importance in the clustering scheme was also investigated. The value of ρ could be based on empirical data considerations. For example, since $0 \leq \rho \leq 1$, the larger the value of ρ the more emphasis is placed on recent years in the data base. In this study the value of ρ to employ was based only on its effect on clustering. Figure 5 shows at what level of the distance scale various numbers of clusters formed as the value of ρ is changed. This Figure suggests that in the vicinity of $\rho = .1$, the points on the distance

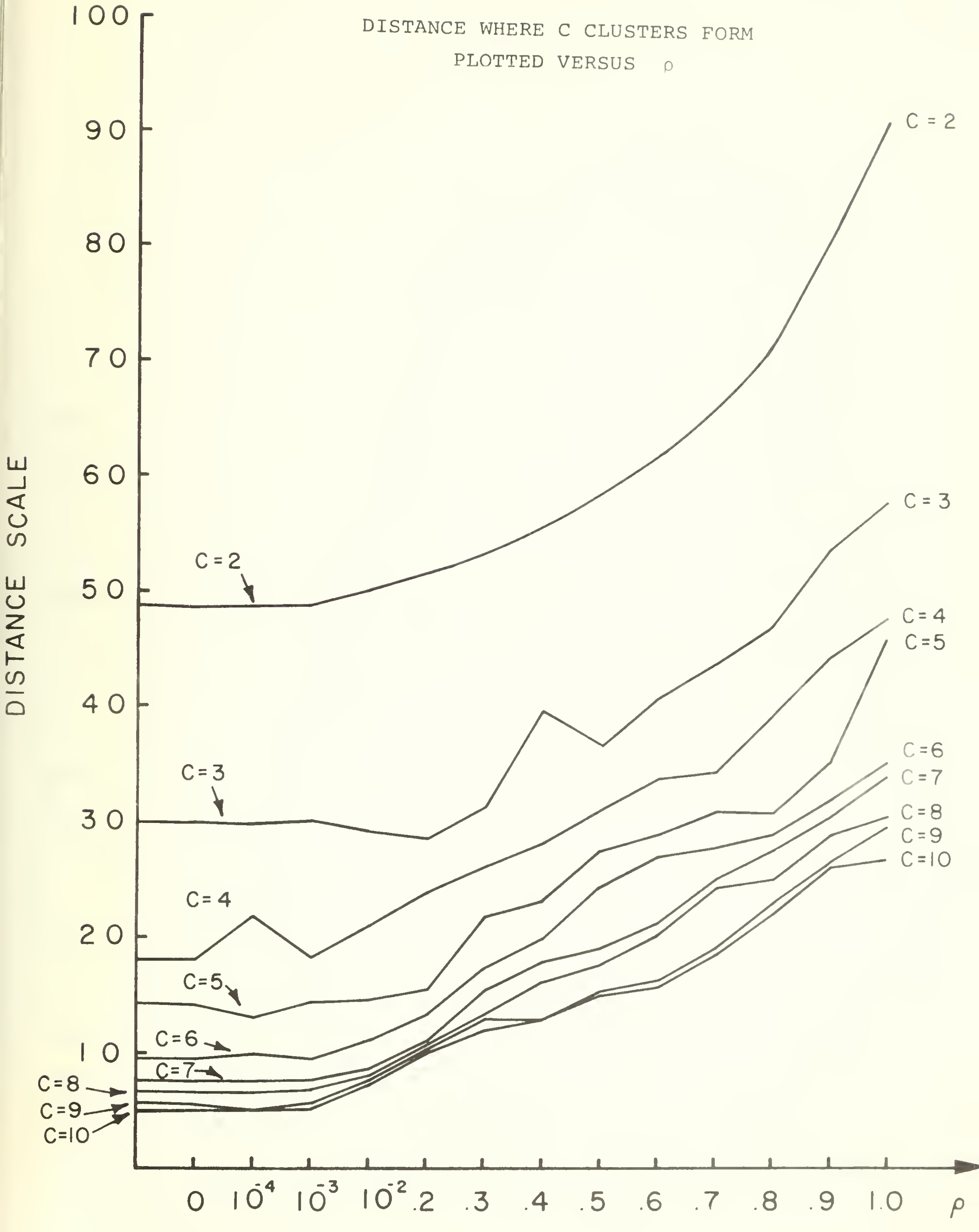

DISTANCE WHERE C CLUSTERS FORM
PLOTTED VERSUS ρ
DISTANCE SCALE
100
90
80
70
60
50
40
30
20
10
0
10^{-4}
10^{-3}
10^{-2}
.2
.3
.4
.5
.6
.7
.8
.9
1.0
ρ
C = 2
C = 3
C = 4
C = 5
C = 6
C = 7
C = 8
C = 9
C = 10

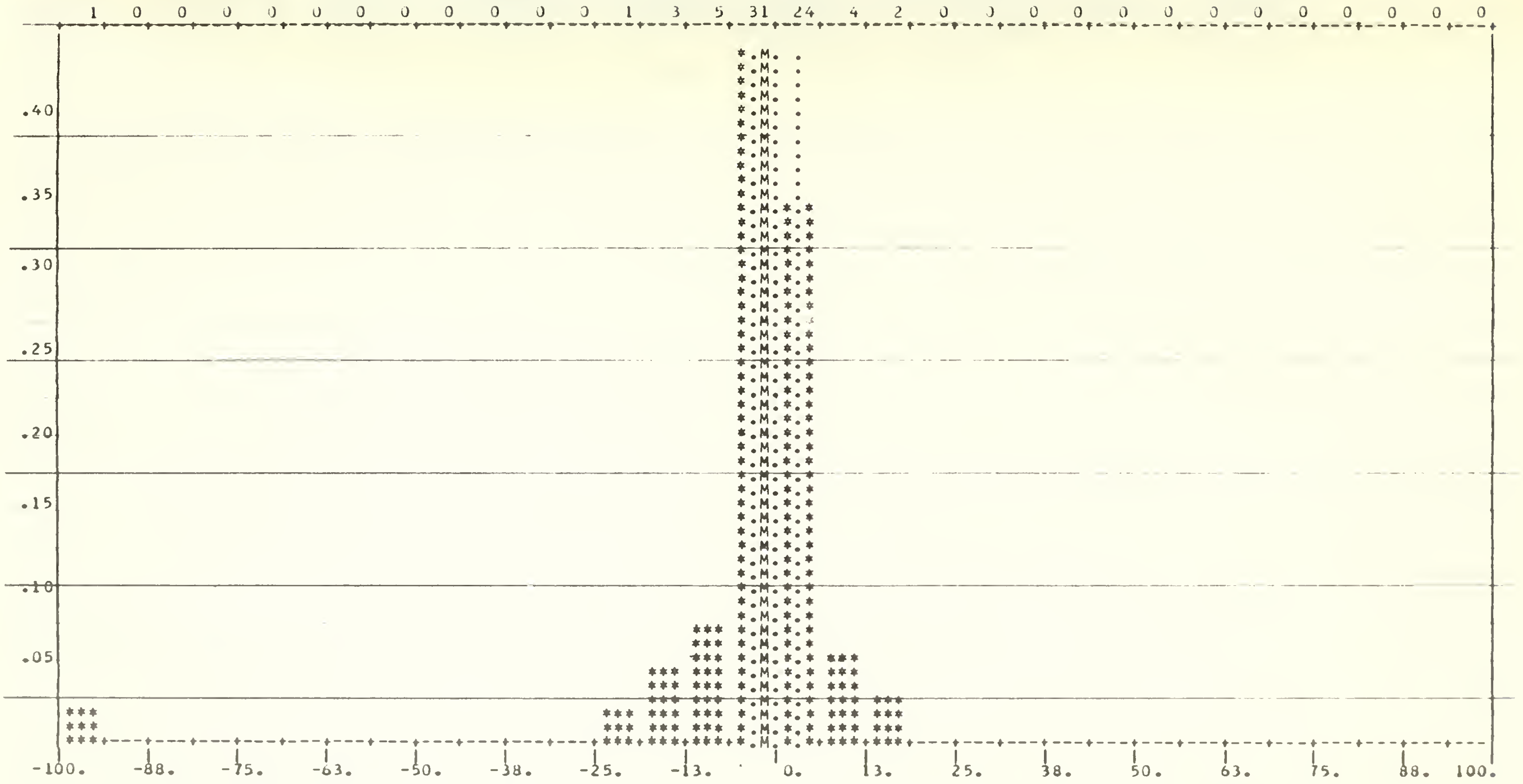
FIGURE 5

scale where clusters form are better separated from each other than is the case for other values of ρ .

The choice of value for α , the parameter that weight past years according to their importance in the prediction scheme, was not specifically investigated. It seemed natural to assume that $\alpha = \rho$. However, there could be convincing arguments for choosing α different from ρ .

Among the types of histograms listed above, item (vi) was the most relevant for the evaluation of clustering. The "difference is absolute normalized errors without and with clustering" measures the relative success of clustering in predicting future losses versus the success of doing that by a comparable traditional method. A large number of ratings having positive values for this measure, especially large positive values, would indicate significant success of clustering. A high percentage of ratings on the negative side would suggest the opposite conclusion. The actual result, however, were not conclusive either way. A typical histogram is shown in Figure 6 for the case is $\rho = .1$ and seven clusters. The mean and median as in most other such histograms are moderately negative, indicating that the clustering was slightly disadvantageous. As more and more clusters are used the histograms become concentrated at the origin which is to be expected, as using many clusters is practically equivalent to no clustering at all. The choice of ρ did not seem to effect this result a great deal, although the choice of $\rho = .5$ appeared to be slightly more favorable to the clustering method. Figure 7

HISTOGRAM OF DIFFERENCES BETWEEN ABSOLUTE NORMALIZED
ERRORS WITHOUT AND WITH CLUSTERING



SCALE FIXED FROM -1.000000E 02 TO 1.000000E 02

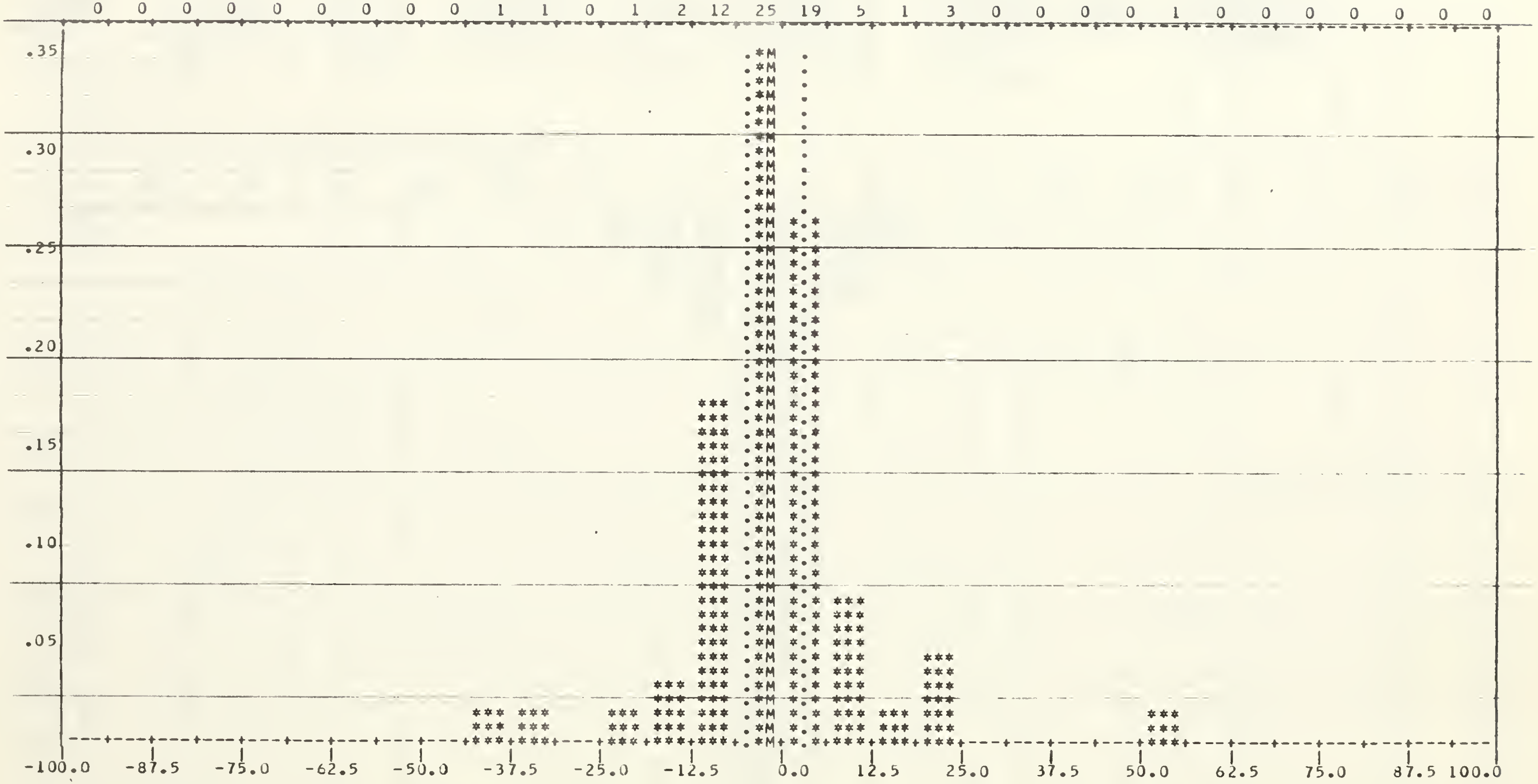
```
CENTRAL TENDENCY              SPREAD                          HIGHER CENTRAL MOMENTS       DISTRIBUTION

MEAN       -1.967990E 00      VARIANCE    1.911424E 02       M3         -1.613011E 04     MINIMUM                       -1.053451E 02
MEDIAN     -7.405243E-01      STD DEV     1.382543E 01       M4          1.707538E 06     .10 QUANTILE                  -7.613548E 00
TRIMEAN    -4.414315E-01      COEF VAR    7.025151E 00       SKEWNESS   -6.103819E 00     .25 QUANTILE  (HINGE)         -3.840118E 00
MIDMEAN    -4.961469E-01      MEAN DEV    5.811298E 00       KURTOSIS    4.373650E 01     .50 QUANTILE  (MEDIAN)        -7.405243E-01
MIDRANGE   -4.481589E 01      RANGE       1.210584E 02       BETA1      -1.545496E 04     .75 QUANTILE  (HINGE)          3.555441E 00
                              MIDSPREAD   7.395559E 00       BETA2       1.614390E 06     .90 QUANTILE                   5.668545E 00
                                                                                          MAXIMUM                        1.571329E 01
```

7 SETS USED. RHO = 0.1000 METRIC = MAXIMUM

FIGURE 7

HISTOGRAM OF DIFFERENCES BETWEEN ABSOLUTE NORMALIZED ERRORS WITHOUT AND WITH CLUSTERING



```
SCALE FIXED FROM -1.000000E 02 TO  1.000000E 02


CENTRAL TENDENCY               SPREAD                         HIGHER CENTRAL MOMENTS      DISTRIBUTION

MEAN      -1.074422E 00        VARIANCE    1.455300E 02       M3         1.127693E 03     MINIMUM                          -4.219064E 01
MEDIAN    -1.385224E 00        STD DEV     1.206358E 01       M4         2.217051E 05     .10 QUANTILE                     -1.190705E 01
TRIMEAN   -1.368165E 00        COEF VAR    1.122797E 01       SKEWNESS   6.423360E-01     .25 QUANTILE (HINGE)             -5.413332E 00
MIDMEAN   -1.289421E 00        MEAN DEV    7.323767E 00       KURTOSIS   7.468164E 00     .50 QUANTILE (MEDIAN)            -1.385224E 00
MIDRANGE   6.092766E 00        RANGE       9.656682E 01       BETA1      1.080492E 03     .75 QUANTILE (HINGE)              2.711118E 00
                               MIDSPREAD   8.124450E 00       BETA2      2.109698E 05     .90 QUANTILE                      7.834747E 00
                                                                                          MAXIMUM                           5.437617E 01
```

7 SETS USED. RHO = 0.5000 METRIC = MAXIMUM

shows the histogram corresponding to the case $\rho = .5$ and seven clusters.

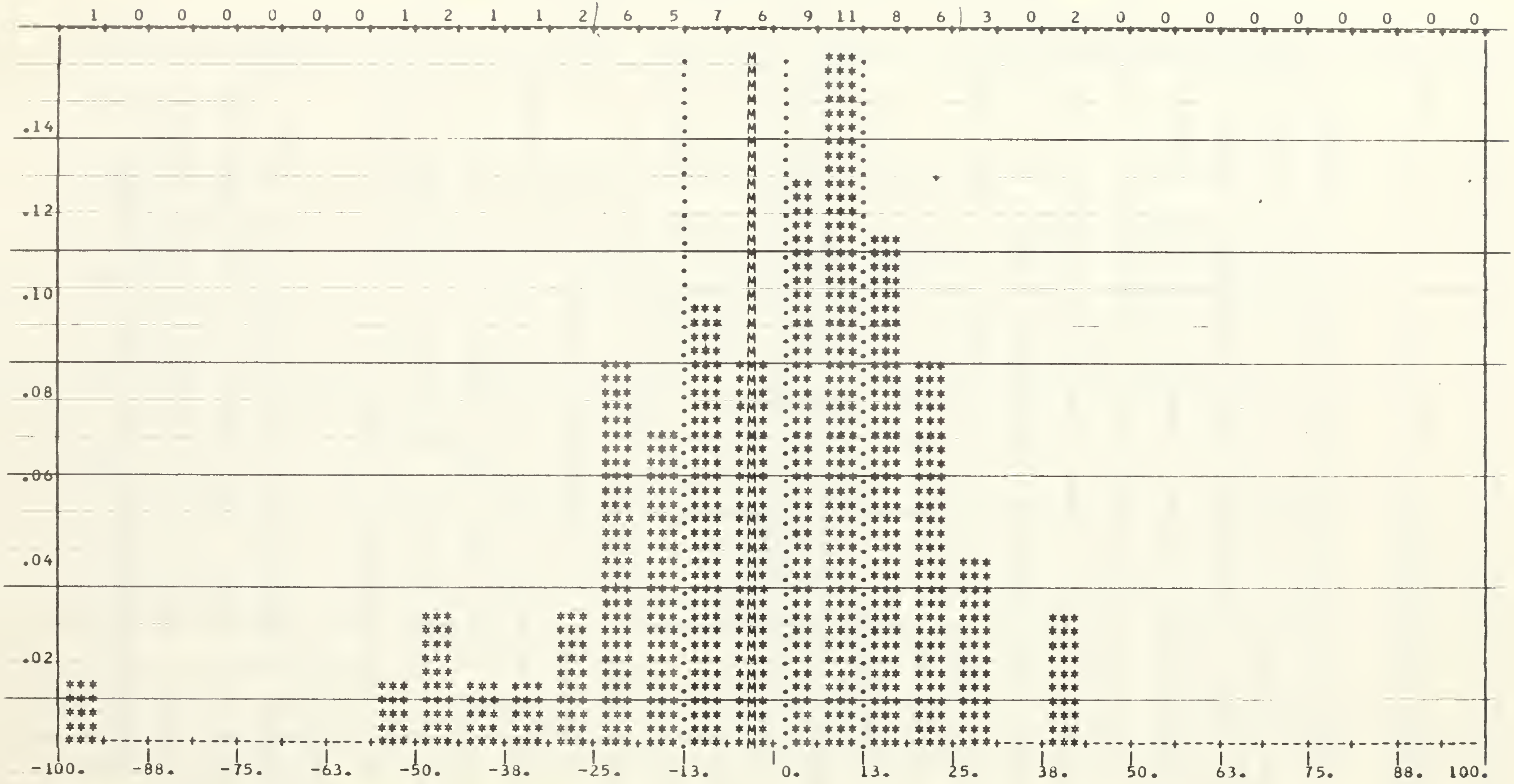
The fact that the clustering method resulted in somewhat bigger (absolute normalized) errors than the standard predicting method does not render clustering totally worthless. Since in comparison the two methods achieve a nearly identical measure of success, the clustering method may have its advantages in shortening the data processing procedures when clustering is used. This may be a more relevant factor when the forcasting technique is not of the simple variety described here, but instead is a more complex one such as used in FAST described in [2], [4] and [5].

The histograms presented above do not show the size of errors made by either the clustering or the standard forcasting method. The histogram presented in Figure 8 exhibits the size of the normalized errors when forcasting by clustering (item (V) above) for the case $\rho = .1$ and seven clusters. The horizontal scale is in percentage. The Figure shows that 58 of the 71 ratings had a less than 25% (positive or negative) error. For one rating the error is shown as -100%. This is due to a rating (Legalman) for which there were zero losses in 1973, while the clustering method forecasted 464. Since the zero loss in 1973 is probably due to a data processing error, this large forcasting error seems forgivable.

The histograms presented here are representive of the many more cases which were tried. The results in every case were essentially the same, namely one of indifference to clustering the data for loss rate prediction. The number of subsets in a

FIGURE 8

NORMALIZED ERROR IN PREDICTION WITH CLUSTERING



```
SCALE FIXED FROM -1.000000E 02 TO  1.000000E 02


CENTRAL TENDENCY              SPREAD                       HIGHER CENTRAL MOMENTS        DISTRIBUTION

MEAN       -3.010119E 00      VARIANCE    8.589490E 02     M3          -8.731725E 04     MINIMUM                          -1.835789E 02
MEDIAN      2.174710E 00      STD DEV     2.930783E 01     M4           1.624445E 07     .10 QUANTILE                     -3.095329E 01
TRIMEAN     1.066062E 00      COEF VAR    9.736434E 00     SKEWNESS -3.468559E 00        .25 QUANTILE (HINGE)             -1.327252E 01
MIDMEAN     1.270288E 00      MEAN DEV    1.793040E 01     KURTOSIS  1.901761E 01        .50 QUANTILE (MEDIAN)             2.174710E 00
MIDRANGE   -7.133551E 01      RANGE       2.244867E 02     BETA1       -8.366244E 04     .75 QUANTILE (HINGE)              1.318735E 01
                              MIDSPREAD   2.645987E 01     BETA2        1.539056E 07     .90 QUANTILE                      2.052130E 01
                                                                                         MAXIMUM                           4.090784E 01
```

7 SETS USED. RHO - 0,1000 METRIC = MAXIMUM

was explored, as well as the choice of the parameters ρ and α. The numerous dendrograms and histograms produced from these experiments remain intact with the authors.

A by-product of this project is the identification of subsets of ratings with common loss behavior. Such a grouping of ratings would for example, sugges guidelines for the application of personnel policy to select groups of ratings. Other applications could be explored as well by simply changing the criterion by which ratings are judged to be close to each other. Then groupings of ratings could quickly and easily be identified, based on another characteristics of behavior besides loss from the service.

REFERENCES

[1] Kendall, M. G., Rank Correlation Methods (2nd Ed.) Hafner Publication Co. 1955.

[2] Structure of FAST (NEE Project) Model. Unpublished Notes, Naval Personnel Research and Development Center, San Diego, California, 1 March, 1972.

[3] Richard O. Duda and Peter E. Hart, Pattern Classification and Scene Analysis, John Wiley & Sons, 1973, pp. 228-236.

[4] FAST. Unpublished Notes, Naval Personnel Research and Development Center, San Diego, California, January 1974.

[5] Boller, Robert L. Design of a Force Structure Model for the Simulation of Personnel Policy. Paper presented at 33rd Military Operations Research Symposium United States Military Academy, West Point, N.Y. June 25-27, 1974.

Initial Distribution List

	No. Copies
Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
Library (Code 2124) Naval Postgraduate School Monterey, California 93940	2
Library (Code 55) Naval Postgraduate School Monterey, California 93940	2
Professor R. W. Butterworth Code 55 Naval Postgraduate School Monterey, California 93940	15
Professor P. R. Milch Code 55 Naval Postgraduate School Monterey, California 93940	15
Mr. Joe Silverman Naval Personnel Research and Development Center San Diego, California 92152	3